Nonlin. Processes Geophys., 14, 73–77, 2007
www.nonlin-processes-geophys.net/14/73/2007/

Nonlinear Processes in Geophysics

Expectation-maximization analysis of spatial time series

K. W. Smith and A. L. Aretxabaleta

Woods Hole Oceanographic Institution, Woods Hole, MA, USA

Received: 16 November 2006 – Revised: 11 January 2007 – Accepted: 17 January 2007 – Published: 1 February 2007

Abstract. Expectation maximization (EM) is used to estimate the parameters of a Gaussian Mixture Model for spatial time series data. The method is presented as an alternative and complement to Empirical Orthogonal Function (EOF) analysis. The resulting weights, associating time points with component distributions, are used to distinguish physical regimes. The method is applied to equatorial Pacific sea surface temperature data from the TAO/TRITON mooring time series. Effectively, the EM algorithm partitions the time series into El Niño, La Niña and normal conditions. The EM method leads to a clearer interpretation of the variability associated with each regime than the basic EOF analysis.

1 Introduction

In recent years, numerical model output, satellite data, and oceanographic observing systems are providing increasing amounts and complexity of spatial time series. Historically, data analyses such as Empirical Orthogonal Function (EOF) analysis have been used to simplify the description of large datasets by separating temporal and spatial variability. Gaussian Mixture Models (GMM) are models that use a mixture of Gaussian distributions to represent a statistical distribution. Expectation-Maximization (EM) can be used for estimating the parameters of a GMM, but applications of EM go beyond that use. In previous studies, EM was used to estimate missing values for oceanographic datasets (Houseago-Stokes and Challenor, 2004; Kondrashov and Ghil, 2006). In the present study, EM is used to estimate the parameters of a GMM yielding a new method to identify regimes in spatial time series and to analyze variability within the regimes.

Cluster analysis is the automatic searching of data in order to group data points into subsets of similar data (Fraley and Raftery, 2002). There are numerous algorithms for accomplishing clustering of multivariate data. We employ an Expectation-Maximization (EM) algorithm because it is simple, and supports a statistical model which has a clear physical interpretation when applied to the analysis of geophysical time series.

In the present study, a subsample of the TAO/TRITON array data, consisting of sea surface temperature (SST) from equatorial Pacific moorings, is used as an application of the EM algorithm. Several important contributions to the understanding of El Niño/Southern Oscillation (ENSO) have relied on data from this array (McPhaden et al., 1998; McPhaden, 1999). The objective of the present application is to evaluate the validity of the EM method as an estimator of ENSO's different regimes. The structure of the article includes a method section, in which we present a summary of EOF analysis, and we introduce the EM method and algorithm; an application section that presents the use of the EM algorithm to moored SST data; and a conclusion section.

2 Methods

2.1 Empirical Orthogonal Function analysis

Empirical Orthogonal Function (EOF) analysis has been broadly used by the entire oceanographic community since it was first introduced. The method and interpretation were explained in several previous studies (Emery and Thomson, 1997). Applications to ENSO time series have been common (Tourre and White, 1995; Tangang et al., 1998; Keeler, 2001).

The basic method can be described as follows: Suppose we have n_d cotemporal time series of length n_t, $\boldsymbol{\psi}(t_m)$ where $\boldsymbol{\psi}$ has length n_d and m runs from 1 to n_t. EOF leads to the decomposition of the time series,

$$\boldsymbol{\psi}(t_m) = \sum_{l=1}^{n_d} \alpha_l(t_m)\boldsymbol{\phi}_l \tag{1}$$

where the $\boldsymbol{\phi}$ form the orthogonal basis consisting of the eigenvectors of the spatial covariance matrix. The spatial

Correspondence to: A. L. Aretxabaleta
(alfredo@whoi.edu)

Published by Copernicus GmbH on behalf of the European Geosciences Union and the American Geophysical Union.

modes, $\boldsymbol{\phi}_l$ describe the spatial structure of variability. For convenience, we assume the index l is sorted by eigenvalue, with $\boldsymbol{\phi}_1$ being the eigenvector corresponding to the largest eigenvalue. It is hoped that these individual modes have some physical interpretation, which can be verified by the time varying amplitudes, $\alpha_l(t_m)$.

2.2 Gaussian mixture models and expectation maximization

A gaussian mixture model (GMM) is a model of a random process whose probability density function (pdf) is the sum of gaussian pdfs. For any point $\psi \in \Re^{n_d}$, the probability can be expressed as,

$$\begin{aligned} p(\boldsymbol{\psi}|\boldsymbol{\mu}^1, \boldsymbol{\mu}^2, ..., \boldsymbol{\mu}^{n_c}, \Sigma^1, \Sigma^2, ..., \Sigma^{n_c}, \tau^1, \tau^2, ..., \tau^{n_c}) \\ = \sum_{k=1}^{n_c} \tau^k \, p(\boldsymbol{\psi}|\boldsymbol{\mu}^k, \Sigma^k) \\ = \sum_{k=1}^{n_c} \tau^k \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\mu}^k)^T[\Sigma^k]^{-1}(\boldsymbol{\psi}-\boldsymbol{\mu}^k)\right)}{\sqrt{(2\pi)^{n_d}|\Sigma^k|}} \end{aligned} \quad (2)$$

where n_c is the number of component distributions, n_d is the length of the data array, τ^k is the probability of component distribution k, and $\boldsymbol{\mu}^k$ and Σ^k are the mean and covariance of the kth component distribution. Here, the component distributions are intended to correspond to different physical regimes.

Expectation Maximization (EM) makes a maximum likelihood estimate of the parameters of a GMM, τ^k, $\boldsymbol{\mu}^k$ and Σ^k, given data $\boldsymbol{\psi}(t_m)$. The EM algorithm produces a fuzzy classification of the data, meaning that a particular time point is not necessarily associated with a single Gaussian distribution, but rather has a probability of arising from any of the GMM's component distribution.

The EM is a two step algorithm with an expectation step and a maximization step. In the expectation step, the likelihood of each data point, $w^k(t_m)$, is computed given the current value of $\boldsymbol{\mu}^k$ and Σ^k,

$$\mathbf{w} = [w^k(t_m)] = \frac{e^{\frac{-1}{2}(\boldsymbol{\psi}(t_m)-\boldsymbol{\mu}^k)^T[\Sigma^k]^{-1}(\boldsymbol{\psi}(t_m)-\boldsymbol{\mu}^k)}}{\sqrt{(2\pi)^{n_d}|\Sigma^k|}} \quad (3)$$

Then, the likelihoods are normalized,

$$w^k(t_m) \rightarrow \frac{w^k(t_m)}{\sum_{k=1}^{n_c} w^k(t_m)} \quad (4)$$

In the maximization step optimal parameters are chosen for the current weights. The frequency of the kth component distribution is computed,

$$n^k = \sum_{m=1}^{n_t} w^k(t_m) \quad (5)$$

Then, the frequencies are normalized,

$$\tau^k = \frac{n^k}{n_t} \quad (6)$$

Finally, the mean and covariance of the kth component distribution are computed,

$$\boldsymbol{\mu}^k = \sum_{m=1}^{n_t} w^k(t_m)\boldsymbol{\psi}(t_m)/n^k \quad (7)$$

$$\Sigma^k = \sum_{m=1}^{n_t} w^k(t_m)(\boldsymbol{\psi}(t_m)-\boldsymbol{\mu}^k)(\boldsymbol{\psi}(t_m)-\boldsymbol{\mu}^k)^T/n^k \quad (8)$$

To begin the algorithm the covariance is taken as the sample covariance and the initial means are randomly assigned from the data. The expectation and maximization steps are repeated until convergence of the w, $\boldsymbol{\mu}$ and Σ is reached. This simple procedure can be shown to converge to a local maximum of the likelihood function (Eq. 2, Fraley and Raftery, 2002). To find the parameters for the global maximum of the likelihood, the EM procedure is repeated several times with different random initial means. The parameters, w^k, $\boldsymbol{\mu}^k$, Σ^k, n^k, corresponding to the highest log-likelihood, describe the GMM.

The w form the basis of our temporal description of the time series, roughly analogous to the temporal amplitudes produced by EOF analysis, $\alpha_l(t_m)$. The means of the component distributions, $\boldsymbol{\mu}^k$, characterize the average behavior of the n_c states of the systems. Although the $\boldsymbol{\mu}^k$ are not orthogonal, the w have convenient properties for decomposing time series.

In practice, we find a tendency for binary behavior, with $w_k(t_m)$=0 or $w_k(t_m)$=1 most often. This clean partitioning of the time domain provids clearer interpretation of regime shifts than the time amplitudes of the EOF analysis. A totally binary w can be enforced in the EM algorithm leading to the so called classification EM (CEM) (Fraley and Raftery, 2002).

In addition to the means of the n_c component distributions, the EM allows us to carry out a *local* EOF analysis, based on the component distributions covariances, Σ^k. Using the eigenvectors of Σ^k, we obtain a new set of orthogonal basis functions which are relevant for times when $w^k(t_m) \simeq 1$. This allows the EOF analysis to be considered independently during different physical regimes. The approach is similar to the method suggested in Tipping and Bishop (1999).

3 Application to TAO/TRITON data

The TAO/TRITON array consist of 70 moorings in the Tropical Pacific Ocean obtaining oceanographic and meteorological data as part of the El Niño/Southern Oscillation (ENSO) Observing System. In the present study, a subsample of the TAO array data, consisting of sea surface temperature (SST) from the equatorial Pacific moorings (including stations along the Equator, and the 2° N and 2° S parallels), is used. The data (Fig. 1, Top) are block averaged between 2° N and 2° S for each longitude (McPhaden, 1999). The resulting

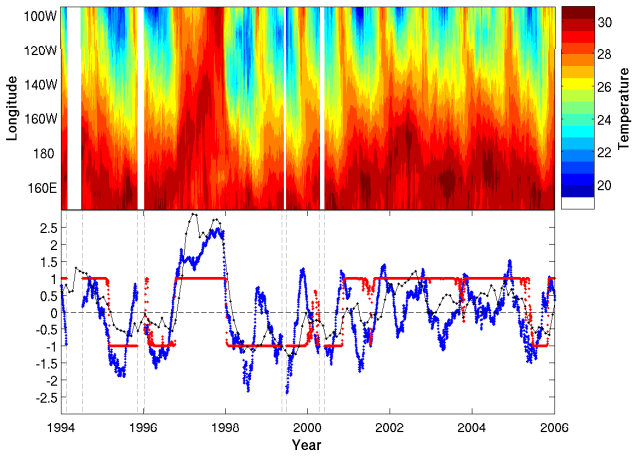

Fig. 1. (Top) Time versus longitude sections of surface temperature (°C) from equatorial moorings. White gaps correspond with periods during which at least one station presented data gaps. (Bottom) Red dots: Probability of Regime A ($w^A(t)$). For representation purposes, the Y-axis has been stretched from the original $0 \leq w^A(t) \leq 1$ to $-1 \leq w^A(t) \leq 1$, such that when $w^A = -1$, $w^B = 1$. Blue dots: Time varying amplitudes of the first EOF ($\boldsymbol{\alpha}_1(t_m)$). Black dotted solid line: Time series of ENSO MEI Index (dimensionless) where positive (*negative*) values correspond with El Niño (*La Niña*) conditions.

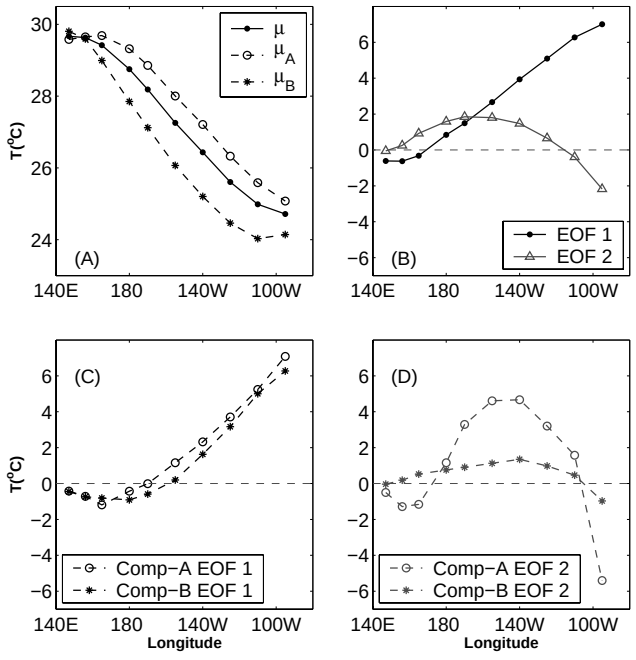

Fig. 2. **(A)** Longitudinal mean temperature for the entire set ($\boldsymbol{\mu}$), Regime A ($\boldsymbol{\mu}_A$), and Regime B ($\boldsymbol{\mu}_B$). **(B)** First and second EOFs ($\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2$) of entire dataset. **(C)** First EOF of Regimes A and B ($\boldsymbol{\phi}_1^A$, $\boldsymbol{\phi}_1^B$). **(D)** Second EOF of Regimes A and B ($\boldsymbol{\phi}_2^A$, $\boldsymbol{\phi}_2^B$).

dataset includes 4271 temporal instances (daily values from 1 June 1994 to 1 June 2006) for each of the 10 longitudinal points considered. For comparison with the EM separation of regimes, the classification of ENSO events followed in the present study (Fig. 1, Bottom) is the NOAA Multivariate ENSO Index (MEI), with positive (*negative*) MEI corresponding to El Niño (*La Niña*) conditions (Wolter and Timlin, 1998).

The gaussian distributions composing the GMM correspond to structures present in the spatial time series of SST. These structures may represent warm/cold states, low variance/high variance states or abnormal (outlier) states.

The application consist of two separate analysis. In the first one, the number of possible regimes is set to two (n_c=2), so the data are separated into either El Niño- or La Niña-like conditions. In the second, an additional possible regime is included (n_c=3) and the data are separated into El Niño, La Niña or normal conditions.

3.1 Two-regime analysis (n_c=2)

The time varying probability of the different regimes, w^k, (red dots on Fig. 1, Bottom) calculated with the EM algorithm shows the method is successful in separating two regimes: Regime A, associated with El Niño conditions ($w^A \simeq 1$); and Regime B associated with La Niña conditions ($w^B \simeq 1$). The 1994–1995, 1997–1998 and 2002–2004 El Niño, and the 1995–1996 and 1999–2000 La Niña events are clearly identified. The method is unable to distinguish between the two regimes during some periods (mid-1996, mid-2001, 2005), which is consistent with normal conditions during that period (MEI∼0). The current transition from La Niña to El Niño (2005–2006) is also recovered by the EM method. The temporal evolution of the amplitudes of the first EOF for the entire dataset ($\boldsymbol{\alpha}_1(t_m)$, blue dots on Fig. 1, Bottom) is able to reproduce the regime changes during most of the record (especially the high signal associated with the 1997–1998 El Niño), but there are several instances in which the EOF evolution registers a change of regime that is not consistent with the MEI index (e.g., during 1999–2000, 2001, and 2002). The correlation between the MEI index and the temporal components for the two methods are similar: the correlation with the temporal amplitudes of the first EOF is r_{eof}=0.71, while the correlation with the EM probability of Regime A is r_{emA}=0.68. When we use the methods as regime estimators (i.e., 1 for MEI>0 and 0 for MEI<0), the EOF estimates the correct regime 75% of the time while the EM estimates correctly 88% of the time. The threshold value (for α_1 and w^A) for regime separation is chosen to optimize the fit to the MEI data.

The spatial (longitudinal) structure of the analyzed data is presented in Fig. 2. The EM method is able to separate different means for the two regimes (Fig. 2A), with a higher mean temperature over the central and eastern Pacific during

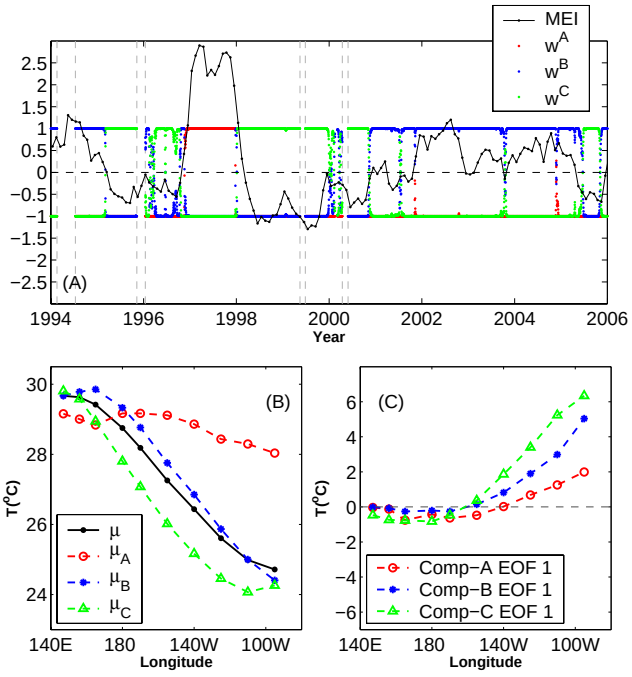

Fig. 3. **(A)** Red dots: Probability of Regime A ($w^A(t)$). Blue dots: Probability of Regime B ($w^B(t)$). Green dots: Probability of Regime C ($w^C(t)$). For representation purposes, the Y-axis has been stretched from the original $0 \leq w^k(t) \leq 1$ to $-1 \leq w^k(t) \leq 1$, such that when $w^A = -1$ and $w^B = -1$, $w^C = 1$. Black dotted solid line: Time series of ENSO MEI Index (dimensionless) where positive (*negative*) values correspond with El Niño (*La Niña*) conditions. **(B)** Longitudinal mean temperature for the entire set ($\boldsymbol{\mu}$), Regime A ($\boldsymbol{\mu}_A$), Regime B ($\boldsymbol{\mu}_B$), and Regime C ($\boldsymbol{\mu}_C$). **(C)** First EOF of Regimes A, B and C ($\boldsymbol{\phi}_1^A, \boldsymbol{\phi}_1^B, \boldsymbol{\phi}_1^C$).

the El Niño regime, and a lower mean during La Niña (1.8°C maximum difference). The means over the western Pacific remain almost identical for all regimes, with Regime A ("El Niño") mean temperature being slightly colder (0.4°C) than the Regime B ("La Niña") mean. Figure 2B shows the spatial structure of the variability of the entire dataset represented by the two first spatial modes ($\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$) of the EOF analysis. The first mode shows the high variability present in the eastern Pacific when compared with the western side. The second mode presents higher variability around the international date line, with the central and eastern Pacific varying in opposite directions.

Both first spatial modes associated with each of the two regimes separated by the EM algorithm (Fig. 2C) are similar to the first EOF of the entire dataset (Fig. 2B, higher variability in the eastern Pacific), with Regime A variability being slightly higher. These results suggests a similar spatial variability structure during both regimes. The second EOF of the separated regimes (Fig. 2D) present significant differences from the second EOF of the entire dataset. While the second EOF remains close to zero for Regime B ("La Niña"), during Regime A ("El Niño") a higher variability is calculated along the central equatorial Pacific and the eastern most station (95° W). The high variability during "El Niño" may be associated with the highly variable intensity and extent of the event in the central Pacific and the occasional rupture of the warm conditions associated with specific wind events in the eastern Pacific (McPhaden, 1999; Belamari et al., 2003). The second EOF has been described as a precursor to the first EOF mode for the Pacific ENSO (Tourre and White, 1995). By separating the regimes using EM, we observed the precursor behavior predominantly during "El Niño", while during "La Niña", the second EOF remains close to zero.

3.2 Three-regime analysis (n_c=3)

The analysis with three regimes (El Niño, La Niña and normal conditions) is presented in this section. The time varying probability of the three different regimes is shown in Fig. 3A): Regime A, associated with strong El Niño conditions ($w^A \simeq 1$); Regime B associated with normal conditions ($w^B \simeq 1$); and Regime C associated with La Niña conditions ($w^C \simeq 1$). Because of its strong temperature signal, the method separates the 1997–1998 El Niño event from the rest of the time series resulting in Regime A. Regime C (La Niña-like conditions) coincides with the 1995–1996 and 1999–2000 La Niña events as well as the the recent La Niña period (2005–2006). Regime B is mostly associated with normal conditions but it includes most of the weak El Niño events as well. The separation between strong "El Niño" conditions (Regime A) and weak "El Niño" and normal conditions (Regime C) is based on the fact that the 1997–1998 period is completely different from any of the other "El Niño" periods.

The EM method is able to separate the different spatial (longitudinal) means for the three regimes (Fig. 3B). The higher mean temperature (up to 4°C warmer) over the central and eastern Pacific during the 1997–1998 El Niño period (Regime A) is significantly different from the other two regimes. The means over the western Pacific remain almost identical for all regimes, with Regime A (strong "El Niño") mean temperature being slightly colder. The structure of the mean temperature for Regime B ("normal conditions") closely resembles the mean temperature of the entire dataset. The mean temperature for Regime C is basically the same as the mean temperature for Regime B in the two-regimes case (Sect. 3.1, Fig. 2A). The first spatial modes associated with each of the regimes separated by the EM algorithm (Fig. 3C) present higher variability in the eastern Pacific, with Regime C variability being higher. Regime A exhibits the smallest variability, which is consistent with the strong temperature signal with small oscillations present during the 1997–1998 "El Niño" period.

4 Considerations and conclusions

Expectation Maximization applied to estimating the parameters of a GMM provides a distinct temporal decomposition relative to EOF analysis (Fig. 1). In the case of the TAO SST data analyzed here, the ENSO signal is recovered cleanly with the EM, while the EOF analysis is ambiguous in relation to the El Niño/La Niña regime separation. The clear separation of the spatial modes achieved by the EM analysis further facilitates the physical interpretation of the data. Therefore, EM is a complement to EOF analysis and an effective regime estimator.

In the present example, the means of the component distributions are the characteristic flavors of the El Niño and La Niña cycles, but this is not always the case in an EM analysis. With some data sets the means may come out to be quite similar, but the covariances may differ in the two distributions (i.e., high variance periods and low variance periods, or periods where correlation between variables changes).

With the n=10 dimensions of our data set and 4271 time points, and assuming two regimes (n_c=2), the EM converged in 50 iterations. When n_d is large, convergence can be hastened by restricting the form of the covariance matrix during the EM. Here, we assumed that the covariance matrix is diagonal, reducing the number of parameters in the GMM from $n_c(1+n_d+n_d(n_d+1)/2)-1$ to $n_c(1+n_d+n_d)-1$. This assumption will almost certainly be necessary for the analysis of global circulation models, satellite data, and other high dimensional spatial time series.

In the present study, the regime separation was evaluated by comparison with an existing ENSO index, but for other applications a similar index may not be available. Using EM to estimate the parameters of a GMM may provide an efficient and reliable method to identify and separate physical regimes in spatial time series. Application of the proposed method to other geophysical data will require careful consideration of the number of regimes to separate.

Acknowledgements. This work was supported by NSF grant DMS-0417845. The TAO/TRITON data were downloaded from the TAO Project Office website (http://www.pmel.noaa.gov/tao/data_deliv/deliv.html). The ENSO MEI values were obtained from K. Wolter NOAA website (http://www.cdc.noaa.gov/people/klaus.wolter/MEI/mei.html). The authors wish to thank L. Fox for her help. We also thank an anonymous reviewer for their helpful comments.

Edited by: G. Zoeller
Reviewed by: one referee

References

Belamari, S., Redelsperger, J.-L., and Pontaud, M.: Dynamic Role of a Westerly Burst in Triggering an Equatorial Pacific Warm Event, J. Climate, 16, 1869–1890, 2003.

Emery, W. J. and Thomson, R. E.: Data Analysis Methods in Physical Oceanography, Pergamon, pp 634, 1997.

Fraley, C. and Raftery, A.: Model Based Clustering, Discriminant Analysis, and Density Estimation., J. American Statistical Assoc., 97, 611–631, 2002.

Houseago-Stokes, R. E. and Challenor, P. G.: Using PPCA to Estimate EOFs in the Presence of Msing Values, J. Atmos. Oceanic Technol., 21, 1471–1480, 2004.

Keeler, W. S.: EOF Representation of the Madden-Julian Oscillation and its Connection with ENSO, J. Climate, 14, 3055–3061, 2001.

Kondrashov, D. and Ghil, M.: Spatio-temporal filling of missing points in geophysical data sets, Nonlin. Processes Geophys., 13, 151–159, 2006, http://www.nonlin-processes-geophys.net/13/151/2006/.

McPhaden, M. J.: Genesis and Evolution of the 1997–1998 El Niño, Science, 283, 950–954, 1999.

McPhaden, M. J., Busalacchi, A. J., Donguy, R. C. J. R., Gage, K. S., Halpern, D., Ji, M., Julian, P., Meyers, G., Mitchum, G. T., Niiler, P. P., Picaut, J., Reynolds, R. W., Smith, N., and Takeuchi, K.: The Tropical Ocean-Global Atmosphere observing system: A decade of progress, J. Geophys. Res., 103, 14 169–14 240, 1998.

Tangang, F. T., Tang, B., Monahan, A. H., and Hsieh, W. W.: Forecasting ENSO Events: A Neural Network-Extended EOF approach, J. Climate, 11, 29–41, 1998.

Tipping, M. E. and Bishop, C. M.: Probabilistic principal component analysis, J. R. Statist. Soc. B, 61, 611–622, 1999.

Tourre, Y. M. and White, W. B.: ENSO Signals in Global Upper-Ocean Temperature, J. Phys. Oceanogr., 25, 1317–1332, 1995.

Wolter, K. and Timlin, M. S.: Measuring the strength of ENSO – how does 1997/98 rank?, Weather, 53, 315–324, 1998.